CLUSTERING WITH SIDE INFORMATION FOR MINING TEXT DATA

[1]Vijayalakshmi P, [2] Dr. P.Marikkannu

[1]PG Scholar, [2]Assistant Professor Department of Information Technology (IT),
Anna University Regional Centre, Coimbatore (India)

ABSTRACT

Side information is available along with text document in several text mining application. They are the different kind of side information such as document provenance information, the link in the document, other non textual attributes which are contained into the document or user access behavior from web logs. Some attributes may contain extremely large amount of information for clustering purpose. Sometimes clustering is more difficult when some of the information is noisy. To design a combination of classical partitioning algorithm with probabilistic model technique to create an effective clustering approach. Then the clustering approach will extend to classification approach for real data set which shows the advantages of previous result.

Keywords: Text Mining, Clustering, Classification.

I INTRODUCTION

Side information is also called meta data which is more available along with text document in several text mining application. Document provenance information, the link in the document, other non textual attributes which are contained onto the document or user access behavior from web logs these are the different kind of side information. When text clustering a problem can comes in some type of application such as web, social networks and also some other digital data. The web contain tremendous amount of text collection in order to create efficient and more scalable algorithms for text mining approach. Web contains meta data for origin of the document. The user access behavior is computing in the formation of web logs. Links in the documents may contain more information for mining process.

Clustering is the task of grouping a set of objects in such a way that objects in the same group, are more similar to each other than to those in other groups called clusters. It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. Types of clustering models: Connectivity models for example hierarchical clustering builds models based on distance connectivity. Centroid models for example the k-means algorithm represents each cluster by a single mean vector. Distribution models clusters are modeled using statistical distributions, such as multivariate normal distributions used by the Expectation- maximization algorithm. Density models for example DBSCAN and OPTICS defines clusters as connected dense regions in the data space. Subspace models in Biclustering (also known as Co-clustering or two-mode-clustering), clusters are modeled with both cluster members and relevant attributes. Group models: some algorithms do not provide a refined model for their results and just provide the grouping information. Graph-based models: a clique, i.e., a subset of nodes in a graph such that every two nodes in the subset are connected by an edge can be considered as a prototypical form of cluster. Relaxations of the complete connectivity requirement (a fraction of the edges can be missing) are known as quasi-cliques.

Text mining, also referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling.

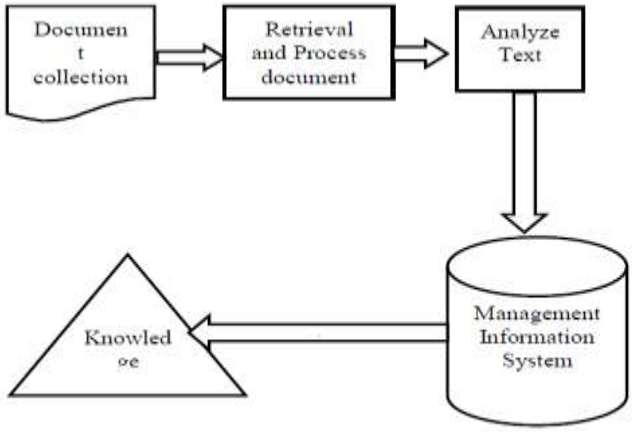


Figure 1: An Example of Text Mining

Classification may refer to categorization, the process in which ideas and objects are recognized, differentiated, and understood. Classification is a data mining (machine learning) technique used to predict group membership for data instances. For example, you may wish to use classification to predict whether the weather on a particular day will be “sunny”, “rainy” or “cloudy”. Popular classification techniques include decision trees and neural networks. There are different types of techniques for classification of the data such as Probabilistic, Naïve Bayes Classifiers, Decision Tree Classifiers and SVM classifiers. The main goal of this paper is to show that the advantages of using side-information extend under a pure clustering process.

II RELATED WORKS

The problem of text clustering was studied in [5][13]. There is a problem in many application such as text crawling, news group filtering and document organization which are requires real time clustering and segmentation of text data records. Using statistical summarization methodology, the problem of text clustering and categorical data streams was solved efficiently. In [37] there is an issue on clustering high dimensional streaming text data. By using combination of an OSKM (online spherical k-means) algorithm with scalable clustering strategy to obtain fast and adaptive clustering of text streams.

Text clustering algorithm for text document is studied in [3][9]. There are many types of algorithm used for text clustering, such as distance based clustering algorithm (like agglomerative), distance based partitioning algorithms (like k-means), Hierarchical Clustering Algorithm and A Hybrid Approach for clustering like scatter/gather technique.

The concept of agglomerative clustering is to successively merge documents into clusters based on their similarities. Virtually the hierarchical clustering algorithms successively merge groups based on the best pairwise similarities between these groups of documents. In a Conceptual manner, the process of agglomerating documents into successively higher levels of clusters creates a cluster hierarchy for which the leaf nodes correspond to their individual documents, and the internal nodes correspond to the merged groups of clusters. A new node is created in this tree corresponding to this larger merged group when two groups are merged.

In [11] Presents the hierarchical data clustering method Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) and it demonstrates that it is especially suitable for large databases. The next method of document clustering is Distance based Partitioning algorithm. Partitioning algorithms are widely used in the database literature in order to efficiently create clusters of objects. K-means clustering algorithm is a partitioning algorithm [3]. It uses a set of k representatives, around which the clusters are built.

In particularly, K-means uses the notion of a centroid, which is the mean or median point of a group. A centroid almost never corresponds to an actual data point. The simplest form of the k-means approach is to start off with a set of k point from the original corpus, and assign documents to these point on the basis of closest similarity. The next iteration, the centroid of the assigned points to each group is used to replace the group in the last iteration. In other words, the new term is defined, so that it is a better central point for this cluster. This approach is continued until convergence. Continuous Clustering and Dynamic Keyword Weighting for Text Documents takes place in [6]. This use the approach to extend K-means algorithm and addition to partitioning the dataset into a given number of clusters, also finds the optimal set of feature weights for each and every clusters. In [13] combination of an efficient online spherical k-means (OSKM) algorithm with an existing scalable clustering strategy to achieved fast and adaptive clustering of text streams.

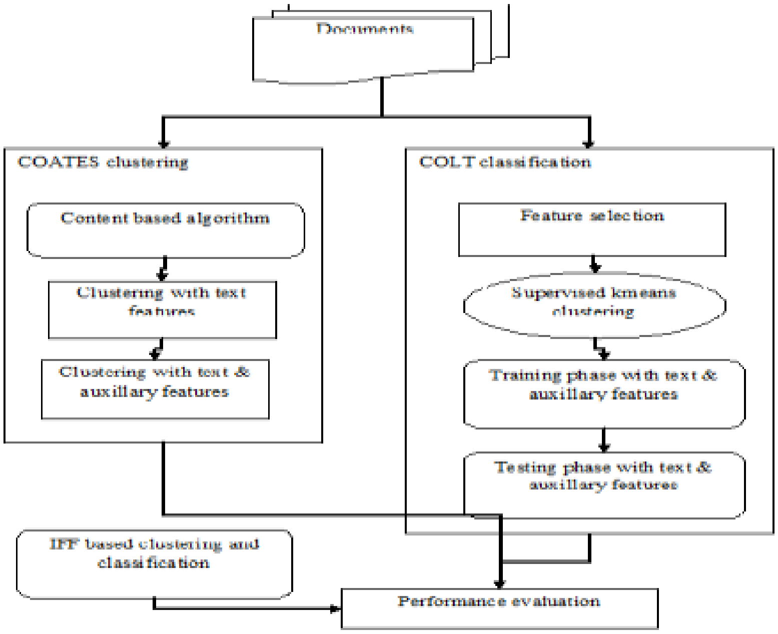
The online spherical k-means algorithm modifies the spherical k-means algorithm (SPKM), using online update (for cluster centroids) based on the well-known Winner-Take- All competitive learning**.** The third type of document clustering is the Hybrid Technique (Scatter-gather technique is the hybrid clustering technique) [5]. An example of the the Scatter/Gather method, which provides a systematic browsing technique with the use of clustered document collection of the document organization. Initially the system scatters the collection of document into a small number of several document groups, or clusters, and presents short summaries of documents to the users. The user selects one or more of the groups for further study based on these summaries. The selected groups are gathered together to form a sub collection documents. Then applies clustering again to scatter the new sub collection into a small number of document groups, which are again presented to the users.

The scatter-gather approach can be used for organized browsing of tremendous amount of document collections, because it creates a natural hierarchy of similar documents. However, these methods are designed for the pure text data clustering , and do not work for in which the text-data is combined with other forms of data. They have some limited work has been done on clustering text in the context of network-based linkage information (like graph mining and algorithms of graph mining in)[2] [10]. A wide variety of techniques have been designed for text classification approaches in [4].

There are different techniques for classification of the data such as Probabilistic and Naïve Bayes Classifiers. Probabilistic classifiers are used to designed an implicit mixture model for generation of the underlying documents. This mixture model is typically assumes that each class is a component of the mixture. Each mixture component is actually a generative model, which provides the probability of sampling to a particular term for that component. Ah another type of classifier is Decision Tree Classifiers which is actually a hierarchical decomposition of the (training data) data space, in which a condition on the attribute value is used in order to divide the data space (training data) hierarchically. The division of the data space is performed recursively in the decision tree until the leaf nodes contain a certain minimum number of records, or some condition on class purity. The most possible class labeling leaf node is used for the purposes of classification[14].

An another type of classifier is SVM classifiers, the main principle of this classifier is to determine separators in the search space which can best separate the different classes. SVM training algorithm builds a model that assigns new examples into one category or other category, making it a non probabilistic binary linear classifier. SVM can efficiently perform a non-linear classification using kernel trick, implicitly mapping their inputs into high dimensional feature spaces. All the work is not applicable to the case of side information attributes. [14] will provide a first approach to using other kinds of attributes in conjunction with text clustering. It show the advantages of using an approach over pure text-based clustering. An approach is especially useful, when the auxiliary information is highly informative and provide an effective guidance in creating coherent clusters. It will also extend the method to the approach of text classification.

III SYSTEM ARCHITECTURE


Documents
COATES clustering
Content based algorithm
Clustering with text features
Clustering with text & auxillary features
COLT classification
Feature selection
Supervised kmeans clustering
Training phase with text & auxillary features
Testing phase with text & auxillary features
IFF based clustering and classification
Performance evaluation

System Architecture Design text are process under COATES algorithm. It has two phases. First phase is content based algorithm used to create cluster of the documents. Second phase is cosine similarity which is used to form centroid of the documents. Gini index used to compute each auxiliary attribute with respect cluster. In COLT algorithm. It have two phases. First phase is Feature selection on text and auxiliary attribute with the use of class labels and gini index the documents. Second phase is cosine similarity which is used to determine the closest cluster of each documents. Gini index used to compute each auxiliary attribute with respect cluster. Then it extend to classification process using COLT classify algorithm which determine the majority class label of the labeled cluster.

IV CLUSTERING TECHNIQUE FOR THE SIDE INFORMATION

In the proposed system, is to show the advantages of using side-information for mining text data extend under a pure clustering process which provides advantages for a wider variety of problems. The COATES Algorithm used for clustering of side information is(COntent and Auxiliary attribute based TExt cluStering algorithm). They has two phases .

1. Initialization

This phase is a lightweight initialization phase in which a Standard text clustering approach is used without any side information. It use the k-means clustering algorithm. In the phase, text document can partitioning data and create a centriod of data to form a cluster. It is based on text information only.

2. Main Phase

This main phase starts off with these initialization phase and iteratively reconstructs these clusters with the use of both text content and auxiliary information . in this step alternating iterations which use the text content and auxiliary attribute information in order to improve the quality of the clustering are performed. These iterations are text content iterations and auxiliary iterations respectively. The combination of the text content iteration and auxiliary iteration is referred to as a major iteration. Each major iteration contains two minor iterations which is corresponding to the auxiliary attributes and text-based methods respectively.

V CLASSIFICATION BASED ON CLUSTERING OF SIDE INFORMATION

In this section to show how to extend the approach to classification. It will extend the earlier clustering approach in order to creates a model which summarizes the class distribution in the data in terms of the clusters. Then, it will show to the use of the summarized model for effective classification. For extension of classification to the problem of clustering , Content and auxiLiary attribute-based Text classification (COLT) algorithm is used for classification of side information .This algorithm uses a supervised clustering approach in order to partition the data into different clusters then the partitioning is used for the purposes of classification. The algorithm works in three phases.

Feature Selection

In the first phase, it uses feature selection to remove the attributes which are not related to the class label. It is performed both for the text attributes and the auxiliary attributes.

Initialization

In this phase, it uses a supervised k means approach in order to perform the initialization with the use of purely text content clustering. The class memberships of the records is in each cluster are pure for supervised initialization. Each cluster only contains records of a particular class when k-means clustering algorithm is modified.

Cluster Training Model Construction

In this phase, a combination of the text content and auxiliary attributes is used for the purposes of creating a cluster based model. During this phase the purity of the clusters is maintained.

VI CONCLUSION

This paper gives the brief introduction about the broad field of document clustering and classification .The techniques which are used for clustering like k-means, hierarchical etc. and classifications like naïve bayes, SVM etc. are discussed. This paper also presented methods for mining text data with the use of side-information. Many forms of text databases contain a large amount of side information or meta-information, which may be used in order to improve the clustering process. In order to design the clustering method, combination of an iterative partitioning technique with a probability estimation process which computes the importance of different kinds of side information takes place. COATES and COLT approach can greatly enhance the quality of text clustering and classification, while maintaining a high level of efficiency.

REFERENCES

[1] Charu C.Aggarwal and Philip S.Yu” On the use of side information for mining the text data”,ieee transaction on knowledge and data engineering,vol.26,no.6,june 2014.

[2] C. C. Aggarwal and P. S. Yu, “A framework for clustering massive text and categorical data streams,” in Proc. SIAM Conf. Data Mining, 2006, pp. 477–481.

[3] D. Cutting, D. Karger, J. Pedersen, and J. Tukey, “Scatter/Gather: A cluster-based approach to browsing large document collections,”in Proc. ACM SIGIR Conf., New York, NY,USA,1992,pp318-329.

[4] M.Steinbach, G. Karypis, and V. Kumar, “A comparison of document clustering techniques,” in Proc. Text Mining Workshop KDD, 2000, pp. 109–110.

[5] C. C. Aggarwal and P. S. Yu, “On text clustering with side information,” in *Proc. IEEE ICDE conf.*, Washington, DC, USA, 2012.

[6] R. Angelova and S. Siersdorfer, “A neighborhood-based approach for clustering of linked document collections,” in *Proc. CIKM Conf.*, New York, NY, USA, 2006, pp. 778–779.

[7] D. Cutting, D. Karger, J. Pedersen, and J. Tukey, “Scatter/Gather: A cluster-based approach to browsing large document collections,” in *Proc. ACM SIGIR Conf.*, New York, NY,USA,1992,pp.318–329.

[8] I.Dhillon, “Co-clustering documents and words using bipartite spectral graph partitioning,” in *Proc. ACM KDD Conf.*, New York, NY, USA, 2001, pp. 269–274.

[9] I. Dhillon, S. Mallela, and D. Modha, “Information-theoretic coclustering,” in *Proc. ACM KDD Conf.*, New York, NY, USA, 2003, pp. 89–98.

[10] P. Domingos and M. J. Pazzani, “On the optimality of the simple Bayesian classifier under zero-one loss,” *Mach. Learn.*, vol. 29, no. 2–3, pp. 103–130, 1997.

[11] T. Zhang, R. Ramakrishnan, and M. Livny, “*BIRCH: An efficient data clustering method for very large databases,*" in Proc. ACM SIGMOD Conf., New York, NY, USA, 1996, pp. 103-114.

[12] S. Zhong, “*Efficient streaming text clustering,*” *Neural Netw.*,vol. 18, no. 5–6, pp. 790–798, 2005.

[13] S. Zhong, “Efficient streaming text clustering,” *Neural Netw.*,vol. 18, no. 5–6, pp. 790–798, 2005.

[14] M. Steinbach, G. Karypis, and V. Kumar, “A comparison of document clustering techniques,” in *Proc. Text Mining Workshop KDD*, 2000, pp. 109–110